The Inner Loop of Collective Human-machine Intelligence

Scott Cheng-Hsin Yang*[1]
scott.cheng.hsin.yang@gmail.com
Tomas Folke[1]
tomas.folke@gmail.com
Patrick Shafto[1,2]
patrick.shafto@rutgers.edu
pshafto@ias.edu
[1]Department of Mathematics and Computer Science
Rutgers University
[2]School of Mathematics
Institute for Advanced Studies

Abstract

With the rise of artificial intelligence (AI) and the desire to ensure that such machines work well with humans, it is essential for AI systems to actively model their human teammates, a capability referred to as Machine Theory of Mind (MToM). In this paper, we introduce the inner loop of human-machine teaming expressed as communication with MToM capability. We present three different approaches to MToM: (1) constructing models of human inference with well-validated psychological theories and empirical measurements; (2) modeling human as a copy of the AI; and (3) incorporating well-documented domain knowledge about human behavior into the above two approaches. We offer a formal language for machine communication and MToM, where each term has a clear mechanistic interpretation. We exemplify the overarching formalism and the specific approaches in two concrete example scenarios. Related work that demonstrates these approaches is highlighted along the way. The formalism, examples, and empirical support provide a holistic picture of the inner loop of human-machine teaming as a foundational building block of collective human-machine intelligence.

Keywords: Artificial Social Intelligence, Human-machine teaming Cognitive Science, Human Computer Interaction

---

Correspondence concerning this article should be addressed to Scott Cheng-Hsin Yang, Department of Mathematics and Computer Science, Rutgers University, 101 Warren Street, Newark, NJ 07102. E-mail: scott.cheng.hsin.yang@gmail.com

1 Introduction

Effective teaming requires that the mental models of teammates be aligned. Humans achieve such alignment by inferring teammates' mental states from observed behavior, and by communicating their own mental states to their teammates. The capacity of humans to model the beliefs, goals, and mental states of others is referred to as Theory of Mind (Frith & Frith, 2005). Similarly, successful human-machine teaming requires both humans and the machines to actively monitor each other and to communicate to ensure alignment in knowledge and goals. Inference and communication are intertwined, forming a loop. Mental modeling of the other agents establishes a shared representation of the team. This representation is updated based on the inference of mental states from the communication signal. The success of communication thus depends on the accuracy of the communicator's inferential model. Repeated communication maintains the alignment of the shared representation across teammates during the length of the task. This nesting of inference and communication is the *inner loop of human-machine teaming* that makes both low-level action coordination and high-level team planning possible.

Human-machine teaming is more challenging than human-human teaming, because fully human teams can rely on substantial shared background knowledge and broadly similar cognitive architecture to support mental model alignment. Because human-machine teams lack some of these common factors, they need to rely on more active inference and communication. How to best communicate the reasoning behind AI decisions to human users is an active area of study and includes the nascent field of explainable AI (XAI) (Gunning & Aha, 2019). XAI has generated a great variety of explanation methods for high-performing machine learning models in various domains. However, these methods still cannot generate explanations guaranteed to be understandable to the humans. This issue occurs largely because these techniques lack accurate models of human inference. Despite these challenges, there are potential benefits to human-machine teams that make them worth the research investment. AI systems are becoming increasingly competent in high-impact domains, including medical (Lundberg et al., 2018), military (Demir et al., 2015), and transportation (Nowak et al., 2019) applications. All of these domains require a human teammate in the loop for accountability, liability, and regulatory reasons. Thus, to fully benefit from the recent improvements in AI, we argue that these systems need to accurately model their human teammates, a capacity referred to as Machine Theory of Mind (MToM) (Rabinowitz et al., 2018).

In this paper, we discuss three avenues that one may go about developing machines that have MToM capabilities. Our contributions in this paper are as follows:

1. We provide a mathematical formulation of the inner loop of human-machine teaming, which supplies a precise language to reason about the core components of human-machine communication (Section 2). See Table 1 for a glossary.

2. We formalize and exemplify a fully psychologically-informed approach to construct the inner loop of human-machine teaming (Section 3).

3. We formalize and exemplify an approach where the machine uses itself to model its human teammates (Section 4). This approach is conceptually similar to how humans use their own beliefs and preferences when modeling others.

4. We describe and exemplify a third approach that augments the previous two approaches by incorporating knowledge about human inference in the domain of interest (Section 5).

The overall formalism and the relationship between the three approaches are depicted in Figure 1. We also discuss the approaches' relative strengths and weaknesses, as well as their implications on collective human-machine intelligence (Section 6).

In a recent survey, Gurney and Pynadath categorize MToM approaches into three broad categories: cognitive architecture, decision theory, and deep learning (Gurney & Pynadath, 2022). Such categorization based on methodology is common in the MToM literature (see, for example, Introduction and/or Related Work in Nguyen and Gonzalez, 2022; Rabinowitz et al., 2018; Raileanu et al., 2018). Our paper advances this categorization by placing MToM approaches in a single framework at the same level of abstraction. In terms of Marr's levels of analysis, the paper suggests that MToM approaches ought to solve the same *computational* problem (Section 2) and that distinctions among them are on the *algorithmic and representation* level (Sections 3–5). The approaches again converge at the *implementation* level, because they are all realized on modern computers. In their review, Gurney and Pynadath also point out that a dire problem in the MToM literature is the absence of "a generally accepted way of comparing the various implementations" of MToM. Our framework suggests that decision-theoretic approaches encompass both the computational and algorithmic levels. Thus, in order to compare these approaches with the cognitive-architecture and deep-learning approaches, one should separate the formulation of the decision objectives from the decision-making process. Once all the approaches are aligned on the same level of abstraction, comparison among them can be carried out more easily, as exemplified and discussed in Section 6. From this perspective, our work offers a more-principled taxonomy of MToM approaches based on their algorithmic and representational choices.

2 Framework

General formulation. We formalize the machine's communication act by the following equation:

$$P_T(x \mid \theta, s) = \frac{P_L(\theta \mid x, s)P(x \mid s)}{\sum_{x' \in \mathcal{X}} P_L(\theta \mid x', s)P(x' \mid s)}. \tag{1}$$

This equation says that the probability of the machine transmitting a particular message, $P_T(x \mid \theta, s)$, is determined by $P_L(\theta \mid x, s)$ — the probability that the message x would lead the human teammate to infer the machine's mental state θ given the current state of the world s. The form of this equation is analogous to Bayesian Teaching, where the transmitter is a teacher (hence the subscript T in $P_T(\cdot)$), and the message receiving teammate is a learner (hence the subscript L in $P_L(\cdot)$) (Yang & Shafto, 2017). The equation can be generalized to cooperative communication if the $P_L(\cdot)$ in turn depends on $P_T(\cdot)$ to form a recursive reasoning loop (Wang et al., 2020). The recursive reasoning can lead to stronger communication and inference if certain assumptions are met, but for this paper, we will focus on the non-recursive version. Below, we define each of the terms in detail.

The *world state* s is a quantity that represents all task-relevant information about the environment and team that has been observed so far. Each teammate can maintain their

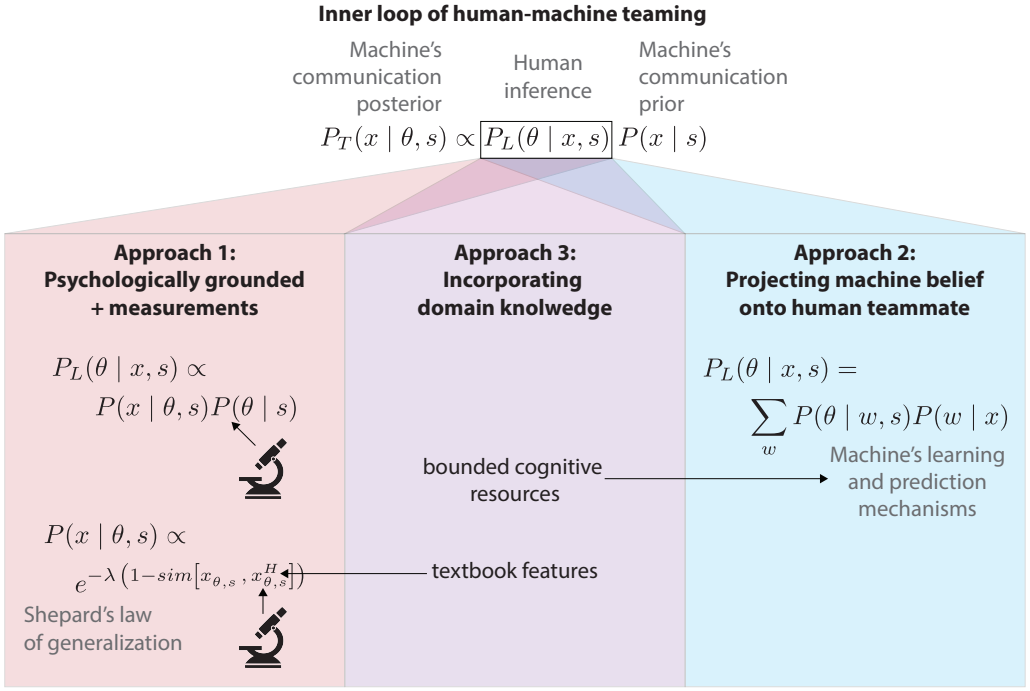

Figure 1

Conceptual overview. The inner loop of human-machine teaming hinges on the machine's communication posterior, which is formed by combining its communication prior and a likelihood that captures human inference (Section 2). We present three approaches to construct this human inference term, the central piece of a Machine Theory of Mind. In Approach 1, human inference is constructed as a posterior distribution with a psychologically grounded likelihood and two measured components (Section 3). In Approach 2, human inference is constructed by projecting the machine's learning and predicting mechanism onto the human teammate (Section 4). In Approach 3, domain knowledge, such as textbook features known to be used by humans and bounded cognitive resources known to limit humans, is incorporated into approaches 1 and 2 to make them more scalable and accurate (Section 5).

own representation of s based on their own knowledge, actions, and observations. However, for effective communication, s should include only the information that is shared and aligned between the transmitter and the receiver. The observable parts of the environment and team often satisfy this constraint. In general, information alignment is related to the establishment and maintenance of common ground. Previous work has shown that communication via a recursion of $P_T(x \mid \theta, s)$ and $P_L(\theta \mid x, s)$ is robust to small perturbations in common ground and admits easy realignment schemes (Wang et al., 2020). In case of large deviation in alignment, a re-establishment of the definition of world states would be necessary.

The *mental state* θ is a quantity that represents some aspects of the machine's inner model. In contrast to the world state, θ is information that is not directly observable by the human teammates. Effective teaming requires that the mental state of any team member be accurately represented by any other team member. Applying this requirement to

human-machine teaming, we constrain θ to carry human friendly semantics, mainly through an intuitive specification of the environment and task. Such a semantic requirement guides the form and complexity of θ to be considered, and suggests that θ should account for the human teammates' individual properties, such as expertise and experience. Effective teaming also requires mental state alignment. Thus, while the scope of the machine's inner model can include models of itself, the task, the environment, the team, and many other things, θ should be focused on certain aspects that would help the human teammate to perform the task given the current world status. Examples of θ include the machine's fine-grained decisions, intent, and goals, as well as higher-level constructs such as a plan.

The *message* x is a collection of data that the machine provides to the human to convey its mental state. Since x is meant for humans, it should be transmitted in a form perceivable to humans. The most common modalities of x are images, texts, speech, actions, as well as sequences and combinations of the above. In reinforcement-learning settings, the machine's actions can serve the dual purpose of performing the task and communicating goals, since humans have an inclination to interpret actions as being goal-oriented (Csibra & Gergely, 2007). A particularly interesting type of x is the machine's explanation of its own actions. The use of explanation allows the task-performing component and the communication component of the machine to be treated separately. That is, the training of the machine can be focused on task performance alone; then, post-hoc explanation techniques can be applied to communicate the machine's mental state. Equation 1 is a framework for the communication component for achieving accurate mental state alignment.

The *communication posterior* $P_T(x \mid \theta, s)$ describes how the machine chooses message x to convey its mental status θ to its human teammates given the current world status s. This posterior probability is determined by the factors on the right hand side of Equation 1, i.e., the teammate's inference posterior $P_L(\theta \mid x, s)$, the communication prior $P(x \mid s)$, and the set of feasible message $\mathcal{X}$. Because the sum in the denominator of Equation 1 is intractable except in the simplest cases, the communication posterior is often estimated through variational or sampling methods that avoid the computation of the denominator (MacKay, 2003). Statistical properties of the communication posterior, such as its entropy and moments, can be diagnostic of the suitability of the message pool $\mathcal{X}$ and the quality of the message itself. For example, a large entropy suggests that no message is uniquely suitable for conveying the mental state of interest. If the communication of this particular mental state is crucial, one might want to consider extending or reconstructing the message pool.

The *inference posterior* $P_L(\theta \mid x, s)$ is the probability that the human teammate will correctly infer the mental state θ intended to be communicated after receiving the message x given the current world state s. Successful communication would cause the inferred θ in the inference posterior to align with the intended θ in the communication posterior. Thus, how much the inference posterior concentrates on the intended mental state is a measure of the goodness of the message. In sections 3–5 we describe three approaches to model human's inference posterior. This term is the centerpiece of our formalism of a Machine Theory of Mind.

The *communication prior* $P(x \mid s)$ is the machine's probability of choosing a particular message x in the current situation s, without considering how that signal would change the teammate's inference about the machine's θ. This term can be used to account for general

information-processing constraints in humans. For example, messages that induce a heavy cognitive load are a priori given a lower probability.

Example scenarios. To ground the terminology, we now describe the components' instantiations in two concrete example scenarios: (1) human prediction of the machine classification and (2) human-machine collaboration involving spatial movements and object manipulation. In the first scenario, the human is asked to predict the machine's classification of an image. This task is useful to evaluate human understanding of machine decision-making. The machine provides an explanatory message to help the human understand its decision process using explainable AI techniques (Yang et al., 2021). The second scenario is typically formalized by a (partially observable) Markov decision process (PO)MDP. In this scenario, the human and machine each controls an avatar. The two avatars are required to coordinate together in order to complete the task efficiently. Task completion generally requires particular coordinated sequences of positional movements and object-handling actions. Examples of such tasks include the Overcooked game (Carroll et al., 2019) and the Door game (Raileanu et al., 2018) in a gridworld, as well as a search-and-rescue task (Huang et al., 2022) and structure-building task (Paleja et al., 2021) in Minecraft.

For the first scenario, s is the image to be classified; θ is the machine's classification of that image; and x can be a visual saliency map intended to show which image regions influence the machine's classification the most, or a small set of images. The communication posterior $P_T(x \mid \theta, s)$ is then the probability of choosing a particular saliency map or image set, and the inference posterior $P_L(\theta \mid x, s)$ is the probability that the human correctly predicts the machine's classification after seeing s and x. The communication prior $P(x \mid s)$ can favor saliency maps that highlight as little area as possible, or image set with a small set size.

For the second scenario, s is the state of the world, including the current positions of the agents and the states of the objects; θ is the machine's goal to move to a certain location or handle a certain object; and x is a real-time view of the machine's actions. The actions, therefore, serve the dual-purpose of performing the task and communicating the machine's mental states. The $P_T(x \mid \theta, s)$ is the machine's probability of taking and showing a certain sequence of actions to communicate its current goal. The $P_L(\theta \mid x, s)$ is the probability that the human teammate would recognize the machine's goal after seeing its actions. The $P(x \mid s)$ can be inversely proportional to the length of the action sequence.

3 Approach 1: Psychologically grounded inference posterior

General formulation. As alluded to in the previous section, the central piece to choosing effective messages from Equation 1 is an accurate model of the inference posterior $P_L(\theta \mid x, s)$, which captures how the human teammate makes inference about the machine's mental state after receiving the message. The cognitive science literature has demonstrated Bayes' rule as a suitable model of human inference in many domains (Chater et al., 2010). Following this literature, we again formulate the human inference posterior via Bayes' rule:

$$P_L(\theta \mid x, s) = \frac{P(x \mid \theta, s)P(\theta \mid s)}{\sum_{\theta' \in \Theta} P(x \mid \theta', s)P(\theta' \mid s)}, \tag{2}$$

On the right hand side, the *inference likelihood* $P(x \mid \theta, s)$ corresponds to the human teammate's belief that the machine would choose message x to convey a mental state θ in

situation s. The *inference prior* $P(\theta \mid s)$ is the probability of an agent entertaining mental state θ in situation s without any communication from other teammates. The denominator is summed over Θ, the set of mental states considered given the current situation s.

The psychology literature suggests that humans tend to project their own beliefs and behaviors onto other agents when forming a theory of mind about them (Buckner & Carroll, 2007). The tendency to project beliefs suggests that the inference prior $P(\theta \mid s)$ can be approximated by the human teammates' own prior about mental state θ. Similarly, the inference likelihood $P(x \mid \theta, s)$ can be modeled by the probability that the human teammate herself would choose the machine-generated x as the message x to be delivered. This assignment of the likelihood probability is a form of generalization [1] from a self-generated message x^H to a machine-generated signal x (Yang et al., 2022). Shepard showed that generalization follows a universal law that decreases monotonically as the psychological distance between two stimuli increases (Shepard, 1987). Borrowing Shepard's celebrated exponential form of generalization, the inference likelihood can be expressed as

$$P(x \mid \theta, s) = \lambda \exp\Big[-\lambda \left(1 - sim\Big[x(\theta, s)\,,\, x^H(\theta, s)\Big]\right)\Big]. \tag{3}$$

Here the $sim[\cdot, \cdot]$ is a similarity function that computes the psychological distance between its two arguments. Its first argument x is the message produced by the machine via Equation 1. The second argument $x^H(\theta, s)$ denotes the message that the human teammate would choose to use to convey mental state θ in situation s. This term is not shown on the left hand side of the equation and is considered something internal to the human teammate. The $(1 - sim[\cdot, \cdot])$ indicates that psychological distance is proportional to dissimilarity. The decrease in generalization is captured by the exponential probability density distribution, where the λ calibrates the rate at which generalization decays with dissimilarity. The final step in fully specifying the inference likelihood is the formulation of the similarity function. Similarity is a well-studied subject in psychology and cognitive science. Notable mathematical formulations include Tversky's ratio model for feature-based similarity (Tversky, 1977) and its continuous extensions (Eelbode et al., 2020).

Scenario on predicting machine classification. To ground this approach in concrete terms, we return to the first example scenario of human prediction of machine classification. In this scenario, a human expert aims to predict a machine's classification θ on an image s given an explanatory saliency map x that highlights image regions critical for the machine's decision. In a recent study, we demonstrated the validity and predictive power of this approach on communicating ResNet-50's classifications on images from ImageNet and Natural Adversarial ImageNet (Yang et al., 2022). Equations 2–3 suggest that we need access to the prior $P(\theta \mid s)$, human's self-generated explanation x^H, as well as a specification of the similarity function in order to construct the inference posterior $P_L(\theta \mid x, s)$. We experimentally probed the inference prior by asking human participants to classify images. We also measured x^H by asking participants to highlight, for each image, critical regions for classifying the image as a particular class. For the similarity function, we used the cosine similarity function, which suits the vector representation of x and x^H obtained and has been

[1] In the cognitive science literature, generalization refers to the probability that a learned response to a stimulus would carry over to be the response to another stimulus; in this case, the stimulus is a message and the response is the act of selecting it.

demonstrated to be psychologically plausible (Sloman & Rips, 1998). The cosine similarity can be expressed as,

$$sim\Big[x, x^H\Big] = \frac{\langle x, x^H \rangle}{\|x\|_2 \|x^H\|_2}, \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors and $\|\cdot\|_2$ denotes the 2-norm. Finally, we measured the actual inference posterior by asking what participants think the AI's classification is, based on the explanation given. These measurements allowed us to compute the fidelity between the model prediction (the inference posterior computed from Equations 2–4) and the human response (the measured inference posterior). The results showed that the model prediction matches the measured $P_L(\theta \mid x, s)$ well (Spearman's $\rho = 0.86$).

The fidelity between model prediction and human response holds over a wide range of inference posterior, suggesting that the model captured the effect of both good and misleading messages. Through a series of ablation studies, we found that removing the inference likelihood from Equation 2, replacing Equation 3 with a non-monotonic decaying likelihood function, and using a less psychologically plausible similarity function[2] all decrease the fidelity significantly. The decrease is strongest when the likelihood is removed or takes on the wrong form, and moderate when the similarity function is ill-informed (see Figure 2D in Yang et al., 2022). These ablation results suggest that Equations 2–3 capture the sufficient core components and their relative importance for modeling human inference of machine mental states given a message. Furthermore, since these equations led to a high fidelity between the model and measured inference posterior over a wide range of probability, using this approach in conjunction with Equation 1 will likely yield good message by virtue of the optimality of Bayes' rule.

Scenario on human-machine collaboration involving action sequences. For this second scenario in a POMDP setting, Nguyen and Gonzalez proposed a similar formalism to compute the inference likelihood $P(x \mid \theta, s)$ (Nguyen & Gonzalez, 2022). Their formalism is based on instance-based learning (IBL) theory, with a memory retrieving mechanism inspired by the well-known cognitive architecture, ACT-R. In their theory, an instance is a tuple of state, action, and outcome/reward. The action maps to the message x, while the state-action-outcome tuple maps to the world state s since all elements of the tuple are observables. The mental state θ is the machine's private knowledge of a high reward object, which is to be inferred from the machine's trajectory. Given instances in the model's memory, the theory computes an unnormalized $P(x \mid \theta, s)$ in three steps, namely, instance activation, instance retrieval, and utility blending. The instances in memory map to x^H; the activation and retrieval functions together map to the similarity calculation in Equation 4; and the blending function map to the computation of $P(x \mid \theta, s)$ in Equation 3. A main result of this approach is that the accuracy of the theory—the model's inference of the acting agent's mental states—matches that of human observers' inference. This is true across a spectrum of task difficulty, indicating the validity of constructing the inference likelihood via psychologically grounded generalization. To apply this IBL approach to generate messages for human teammates, one would insert measurements of human trajectories and outcomes into the model's memory and then compute Equations 1–3 in reverse order. Given that the

[2] A pixel-by-pixel L1 norm distance.

inference likelihood captures human inference of machine mental states well, the message sampled from the machine's communication posterior should also be effective by virtue of the optimality of Bayes' rule.

Ho et al. provided further evidence that this approach is effective by showing the message generated from Equations 1–3 match messages generated by human teachers (Ho et al., 2016). In their work, the goal of the machine is to demonstrate to an observer a hidden goal grid (the θ) among several target grids by positional movements. They mapped Equations 1–3 to the probability of a teacher's demonstration, the probability of an observer's inference of goal grid, and the probability of taking certain actions under a given policy, respectively. Because the task environment is simple, the policy that a human would use in this case is the same as the optimal policy that a machine would obtain by standard reinforcement learning. In other words, the optimal policy plays the role of x^H in Equation 3. They first showed that while simple task-performing actions are benevolent to ambiguous paths as long as the goal is achieved, intentional demonstrations favor paths that allow non-ambiguous inference of goal grid from earlier parts of the trajectory. Then, they showed that the demonstrations generated from their formalism match what human demonstrators do for teaching the goal grid. This match further validates Approach 1 because human-human communication is arguably the current gold standard.

Simplifying state space. The main difficulty this approach encounters in the POMDP setting is that the number of world states, mental states, and messages one might consider may be very large. Consequently, naive measurement of the inference prior $P(\theta \mid s)$ and the x^H may not be feasible. To make the measurement space more manageable, one should constrain the space of s, θ, and x as much as possible while not compromising on the effectiveness of the output message. The number of world states can be reduced through abstraction—a distillation of a large number of observable spatio-temporal world states into a small number of latent states that are in some sense equivalent (Abel, 2022; Jong et al., 2008; Li et al., 2006). A reduction in the number of world states often leads to a reduction in the number of mental states. This is because whether a mental state should be considered largely depends on the situation encountered, as implied by the dependence of the inference prior $P(\theta \mid s)$ on s. The effective number of mental states can be further trimmed by assigning non-zero prior probability to only mental states that are critical for performing the task (Ho et al., 2022) and relevant according to a situational awareness analysis (Endsley, 2015). Moreover, the inference of θ from x and vice versa ought to be as straight-forward as possible to avoid unnecessary ambiguity. Theories of rational speech-act model (Frank & Goodman, 2012) and optimal cooperative inference (Yang et al., 2018) suggest constraints on the inference likelihood that guarantee accurate and efficient referencing of θ from x. Specifically, if the global structure of the inference likelihood on the sets of mental states θ and messages x satisfies certain hierarchical and sparsity constraints, each x will point to only one θ with high probability, leading to effective communication of particular θ.

4 Approach 2: Inference posterior based on projection of machine beliefs

General formulation. In the second approach, the machine projects its own learning mechanism onto the human teammate and selects the message x as if the human would learn from x in the same way that it itself does. Mathematically, the inference

posterior takes on a new form:

$$P_L(\theta \mid x, s) = \sum_w P(\theta \mid w, s) P(w \mid x). \tag{5}$$

Here w denotes the set of parameters that fully specifies the machine. The $P(w \mid x)$ describes the machine's learning mechanism, that is, how w is updated given the training data x. The $P(\theta \mid w, s)$ corresponds to the machine's prediction mechanism, that is, how the machine's mental state changes when encountering s given the particular model specification w. The sum over w is the formal Bayesian treatment to marginalize out uncertainties in the model specification.

The idea behind this approach is to approximate human inference by combining the machine's learning and prediction mechanisms. From a theory-of-mind perspective, the machine assumes that the human is a copy of itself, but has not been exposed to the world state s yet. The message x can thus be thought of as a pedagogical summary of s that would lead the less-experienced copy of the machine to arrive at the mental state θ, which the more-experienced copy of the machine developed after encountering s. The machine learning literature offers a variety of frameworks to specify $p(w \mid x)$ and to generate the final x. To name a few, these frameworks include machine teaching (X. Zhu et al., 2018), influence function (Koh & Liang, 2017), and coreset (Bachem et al., 2017). The common thread across these frameworks is the selection of additional training data (the message x) to help the recipient (the model of the human teammate) arrive at the desired state (θ). From the angle of training machines, this idea shares the flavor of fine-tuning and transfer learning (Iman et al., 2022).

Contrasting Equations 3–4 with Equation 5, we see that the second approach replaces the Bayes' rule and the psychologically grounded likelihood with the machine's own learning and prediction mechanism. The justification for such replacement hinges on the machine's ability to perform the task well, which can be traced back to the training of the machine. We adopt the practical assumption that past world states (or data of the same form) are used as training data. Under this assumption, high machine performance implies: (i) that the training process can distill useful information from the training data into the machine; (ii) that the machine's mapping — characterized by w — from world state to mental state is meaningful; and (iii) that the machine's representation of θ is useful for performing the task. The world state s is typically constructed or curated by humans and hence human interpretable; thus, this approach is most fruitful if the message x is also constrained to have a similar form to s. Constraining x in this way will ensure that the message is human interpretable. Furthermore, because the training based on the training data similar to s led to meaningful w, messages constrained to have the form of s will suit the machine's learning mechanisms well. Below, we return to the two example scenarios to exemplify this second approach in more concrete terms.

Scenario on predicting machine classification. In a previous work, we studied the effectiveness of this approach on the prediction-of-classification task (Yang et al., 2021). The current world state s is an image to be classified. The θ is the classification of that image predicted by a ResNet50 model trained on the training data set of ImageNet. The machine used to approximate the inference of the human agent is a ResNet50-PLDA model, where the final fully connected layer of the RestNet50 model is replaced by a probabilistic

linear discriminant analysis (PLDA) model. In other words, the initial w includes all the weights of the neural network up to the last layer plus the parameters of the new PLDA layer. In the spirit of transfer-learning style inference, we fix the neural network's weights but allow the parameters of the PLDA to change. Thus, the $P(w \mid x)$ in Equation 5 corresponds to the training of the PLDA. The predictive distribution $P(\theta \mid w, s)$ characterizes the prediction of the entire ResNet50-PLDA model. The message x takes the form of a small subset of images to match the form of s. The message x is selected by maximizing the inference posterior $P_L(\theta \mid x, s)$ computed using Equation 5. That is, this message x is a small set of additional training images that would lead the approximate human inference model to the desired θ.

Empirical results show that the messages produced following Equations 5 and 1 have a positive effect in shifting people's mental state to the targeted θ. Specifically, people's prediction of the machine's classification after receiving x correlate with the inference posterior computed by Equation 5. Further analysis shows that the effectiveness of the message x positively correlates with the machine's category-level accuracy and human's familiarity with the classes in question. These modulation effects suggest that the message is most helpful when human and machine are aligned in background knowledge as captured by the machine's performance and human's familiarity on categories, and the communication effectiveness deteriorates as the alignment weakens. Furthermore, the message is effective when the machine's classification matches ground truth, but is ineffective otherwise. The machine's misclassification exposes where the representation of world state s differs between human and machine; thus, the alignment in the representation of s between human and machine is also a crucial factor for the success of this approach.

Scenario on human-machine collaboration involving action sequences. Because the inference posterior in this approach is based on the machine's representation, it is most convenient to express s, θ, x, and w by quantities in the (PO)MDP, which representation is shared by both the machine and the human. Thus, we describe the world state s by the locations of the agents and the states of the object, formalize the mental states θ (the machine's goals) by a pair of current and future world states with a horizon, and restrict the modality of the message x to the modalities of the training data. The machine's original w can be obtained by standard multi-agent reinforcement-learning training techniques such as self-play. With the above specification, the message is a short sequence of state-action pairs (or a short trajectory) aimed to communicate the machine's interim goal, for example, to go and grab a certain object. Equation 5 says that the message is first processed by the term $P(w \mid x)$, indicating that the agent would update its parameters w from the short trajectory x. The update can be interpreted as a form of learning from demonstration and thus admits techniques from imitation learning (Zheng et al., 2021). Once the parameters are updated, the prediction term $P(\theta \mid w, s)$ can be derived from state-action value functions along planning trajectories that lead to the desired world states specified by θ.

The machine learning literature offers qualitative insights on how well this approach works. Rabinowitz et al. proposed a Theory of Mind neural network (ToMnet) to infer an agent's mental states from the agent's actions (Rabinowitz et al., 2018). As such, the function of a ToMnet is analogous to our inference posterior. ToMnet's architecture is not based on known cognitive architecture but inspired by a machine learning technique called meta-learning, making it an approach based on the projection of machine beliefs. The

authors show that ToMnet can predict the goal-directed actions of random, algorithmic, and deep RL agents. The coverage of a wide range of agents suggests that the approach has the capacity to model human, and the successful prediction of goal-directed actions implies accurate inference of the acting agent's mental state. Again, by the virtue of Bayes' rule (Equation 1), an accurate inference posterior will likely lead to effective communication.

Raileanu et al. also proposed a neural network for inferring mental states (Raileanu et al., 2018). Their network's architecture is inspired by Self-Other-Modeling (SOM), that is, modeling other's action by asking what my mental state would be if I had acted as the other player had. The Self-Other-Modeling philosophy is analogous to our idea that the machine makes inference about the human by assuming the human as a copy of itself in Approach 2. The authors showed that two agents equipped with the SOM architecture can infer each other's mental states and achieve decent reward in several cooperative games. The reward is not as high as agents who have access to each other's true mental states, but is considerable higher than agents who do not infer the mental states of the other agent. While communication is not intentional in this scenario, the decency with which cooperative tasks are completed suggests that intentional communication of mental states will also be effective to certain extents. In a speaker-listener scenario where communication is explicit, Zhu et al. showed the same qualitative result: a speaker agent equipped with MToM gives better direction than one without MToM, but performs worse than an agent having access to the actual mental states (H. Zhu et al., 2021).

A major deficiency of MToM agent in Approach 2 comes from the inherent difference between human planning and the policy derived from RL training. Carroll et al. confirmed that collaborative performance depends on mental model differences (Carroll et al., 2019). They showed that machines trained with self-play or population-based techniques can coordinate well with themselves but not humans; in contrast, machines that learned from human demonstrations can coordinate better with humans than machines that did not.

5 Approach 3: Domain-knowledge augmented inference posterior

General idea. The general idea of the third approach is to incorporate knowledge of human behavior in the domain of interest into the construction of the inference posterior. This approach relaxes Approach 1 by replacing quantities that are difficult to measure in Equations 2–3 with models built from knowledge about human cognition in the task of interest. Approach 3 tightens Approach 2 by infusing known human biases into the machine to make the machine more human-like. In other words, Approach 3 is a middle-ground approach between Approach 1, which is grounded in human psychology and based on measurements, and Approach 2, which relies heavily on the machine's training and prediction. To illustrate these ideas, we return to the example scenarios.

Scenario on predicting machine classification. Here, we illustrate relaxing the measurement requirement of Approach 1 by using well-established expert knowledge. For the prediction-of-classification task, the prior $P(\theta \mid s)$ can often be measured by asking the human expert's classification of each image s, but measurements of x^H are typically much more challenging. For example, while dermatologists provides their diagnoses on images of skin lesions as part of their routine, they usually do not have the time to engage in a detailed documentation of their self-generated explanation. Fortunately, for many high-stake classification problems, the features that human experts pay attention to are

well-documented. These well-documented features, which we refer to as *textbook features*, can be algorithmically encoded to generate approximate x^H. Using the skin lesion example, medical textbooks teach dermatologists in training to diagnose melanoma using the ABCDE rule (Abbasi et al., 2004), referring to five particular image features. All of these features can be extracted with image processing and machine vision techniques (Smaoui & Bessassi, 2013), and the extracted features can be used to approximate the self-generated x^H. Once the x^H's have been encoded, the machine can trace backwards from Equation 4 to Equation 1 to produce a desirable message x for its classification θ on an image s. The validity of this approach rests on how good of a proxy the encoded textbook features are to the features human agents use. A simple assessment of the quality of the encoded features is to use these features in a simple linear model to check whether they carry sufficient discriminant power for the classification task. In a recent work we showed that a logistic regression model based on the ABC features (Asymmetry, Border, and Color) performs as well as dermatologists with 3 to 5 years of experience in diagnosing melanoma (Bokadia et al., 2022). This result suggests that Approach 3 can produce reasonable inference posterior, and in turn, legitimate messages.

Scenario on human-machine collaboration involving action sequences. Now we illustrate tightening Approach 2 by incorporating known human factors into the machine. A general framework to do so for reinforcement-learning agents is expected utility theory with bounded resource, also known as resource rationality or bounded optimality (Ho & Griffiths, 2022). This framework keeps the generality of the reinforcement-learning training but makes the trained machine more humanistic by constraining the machine to have limited cognitive resources in the way that humans do. For example, compared to machines, humans have very limited computational resources to plan ahead. One way that humans deal with this limitation is state abstraction. While there exist many techniques for state abstraction as mentioned in Section 3, the machine will be more aligned with the human teammate if it uses the kind of abstraction that humans use. Ho et al. showed that humans construct a simplified representation of the environment for efficient planning, called a task construal, by considering only cause-effect relationships in the environment (Ho et al., 2022). In particular, using a gridworld maze, they demonstrated that humans pay more attention to critical obstacles—that is, obstacles that are highly relevant for path planning —than to irrelevant obstacles, independent of how far the critical obstacles are from the optimal path. Since a standard RL agent would represent and account for all obstacles by default, incorporating this specific human factor of task construal would highlight the mental states that humans are biased to pay attention to and help the machine produce better messages. Other ways to incorporate human factors into the machine includes supervised-learning of policy networks from expert behavior (Silver et al., 2016) and reward shaping, either by direct reward encoding or inverse reinforcement learning.

6 Discussion

Having introduced the three approaches to construct the inference posterior, we now discuss the approaches' relative strengths and weaknesses along four dimensions—accuracy, transparency, personalization, and scalability—and summarize the conclusions in Table 2.

Accuracy. Here accuracy refers to the accuracy by which the machine can infer human mental states as captured by the inference posterior. An accurate inference posterior

leads to an effective message by virtue of the optimality of Bayes' rule. We expect Approach 1 to be the most accurate, followed by Approach 3, and Approach 2 to be the least accurate. The main factor that determines accuracy is the level of alignment between the machine's model of human inference and human's actual inference. In the scenario on predicting machine classification, our previous work shows that Approach 1 captures human response nearly perfectly (Yang et al., 2022) while Approach 2 gets the trend but not the magnitude of the responses (Yang et al., 2021). We expect the accuracy of Approach 3 to be in between these two approaches, because textbook features would likely align with human reasoning better than black-box models but worse than detailed measurements of latent mental states.

In addition to misalignment of mental models, the more complex scenario of human-machine collaboration includes another factor that can reduce accuracy: as the environment and task become complex, a message may be equally effective to convey multiple mental states, creating ambiguity and uncertainty in the inference posterior. The issue of mental state misalignment is illustrated by the result that machines can better collaborate with humans when trained on human behavior (Carroll et al., 2019). The second issue of uncertainty in inference is exemplified by the results that modelling another agent as itself is not as good as having access to the actual mental state of the modelled agent (Raileanu et al., 2018; H. Zhu et al., 2021). Based on these two factors, we expect the ranking of Approaches 1–3 in the human-machine collaboration scenario to be the same as that in the prediction-of-classification scenario. In principle, Approach 1 eliminates the mental model misalignment by measuring the relevant, latent mental states. In practice, the feasibility of Approach 1 implies the world has been simplified enough to make the measurement space manageable, which in turn reduces the probability of encountering ambiguous mapping between messages and mental states. Similarly, Approach 3 promotes mental model alignment and a more-compact description of the world by restricting the machine with resource and inference constraints known to affect humans (Ho et al., 2022). In contrast, Approach 2 by definition lacks explicit mental model alignment, and often does not deal with the complexity-induced uncertainty explicitly.

Transparency. Transparency refers to the ease with which humans can understand how and why the machine generates the messages it generates. As such, the notion of transparency includes concepts such as interpretability, explainability, and the complexity of the machine's inner workings. The transparency ranking of the approaches largely follows the ranking of accuracy because mental model alignment is also an important determinant of transparency. In the language of our framework, it is easier for humans to reason about how inferences can be derived from measurements of human latent mental states x^H (Approach 1) and textbook features (Approach 3) than from the machine's parameters w (Approach 2).

Personalization. Personalization refers to the ability to provide nuanced messages suitable for a particular teammate. Approach 1 is inherently more personalized than Approaches 3, as individualized measurements of latent mental state are more personalized than textbook features. Approach 2 can achieve personalization to some extent via meta-learning: the idea is to build up a rich prior among θ, s, and x by training the machine on a good variety of scenarios so that the machine could quickly adapt to new scenarios from just a few novel training data (Rabinowitz et al., 2018; H. Zhu et al., 2021). From the perspective of personalization, the different scenarios refer to different human teammates, each of whom may have their way of reasoning.

Scalability. We mainly estimate scalability by the amount of work required to obtain the training data needed to implement each of the three approaches. Approach 1 requires the measurement of the inference prior $P(\theta \mid s)$ as well as the self-generated message x^H on an instance-by-instance basis for each θ, s, and human teammate. Thus, Approach 1 is most suitable for simplified settings where the numbers of world states and mental states are small. Approach 2 requires many fewer measurements. For machine classifiers, pairs of s and the most likely θ are sufficient to ensure decent task performance. For autonomous machines, one only needs to specify the MDP or POMDP environment. These measurements and specifications are much less labor-intensive compared to Approach 1; thus, Approach 2 is suitable for more complex tasks. With Approach 3, the labor of measuring x^H is replaced by the labor of finding and encoding the knowledge, which is less personalized but more scalable. In augmenting Approach 2, Approach 3 should either supply expert demonstrations or re-design the environment and training. Overall, Approach 3 is also suitable for complex tasks if algorithmic encoding of knowledge is feasible.

Building blocks. Communication is a central component of collective *human* intelligence. Task coordination, knowledge accumulation, and even cultural evolution all depend on effective communication. We argue that communication should also be a central component of collective *human-machine* intelligence. This paper introduces the *inner loop of human-machine teaming*, which focuses on how machines can communicate with humans. This inner loop features the dependence of communication on inference, a form of Machine Theory of Mind. Formalizing the inner loop yields a formal definition of machine communication and formal models of human inference that are implementable by the machine. In addition to presenting the formalism, this paper also provides examples of formal models of human inference across a wide spectrum of representation. Lastly, a related building block is the machine's capability to understand natural modalities of human communication so that humans can communicate with the machine with ease (Chen et al., 2018). These three building blocks — Machine Theory of Mind, machine's communication of its "mental" states, and machine understanding of human communication — form a basic machine capable of communicating with humans, and hence, a foundation of collective human-machine intelligence.

In conclusion, our main contribution is an exemplified formulation of the inner loop of human-machine intelligence that features the mutual dependence between human-machine communication and Machine Theory of Mind (MToM). This work offers a more-principled taxonomy of MToM approaches and shows how different representation of MToM could be constructed. The framework positions existing approaches in the MToM literature from both the machine-learning and cognitive-science communities under one coherent framework on the same level of abstraction. This principled repositioning facilitates comparison among the approaches and makes explicit the connection between cognitive and machine-learning models of MToM. We consider this connection as an initial but necessary step towards the emergence of collective human-machine intelligence. A future direction of research is to extend the scope of world states, mental states, and inference models in this framework to include more aspects of practical human-machine intelligence, including practical factors such as natural language communication and emotion factors such as affect.

References

Abbasi, N. R., Shaw, H. M., Rigel, D. S., Friedman, R. J., McCarthy, W. H., Osman, I., Kopf, A. W., & Polsky, D. (2004). Early diagnosis of cutaneous melanoma: Revisiting the abcd criteria. *Jama*, *292*(22), 2771–2776.

Abel, D. (2022). A theory of abstraction in reinforcement learning. *arXiv preprint arXiv:2203.00397*.

Bachem, O., Lucic, M., & Krause, A. (2017). Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*.

Bokadia, H., Yang, S. C.-H., Li, Z., Folke, T., & Shafto, P. (2022). Evaluating perceptual and semantic interpretability of saliency methods: A case study of melanoma. *Applied AI Letters*, e77.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in cognitive sciences*, *11*(2), 49–57.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., & Dragan, A. (2019). On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, *32*.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811–823.

Chen, Y.-N., Celikyilmaz, A., & Hakkani-Tur, D. (2018). Deep learning for dialogue systems. *Proceedings of the 27th international conference on computational linguistics: Tutorial abstracts*, 25–31.

Csibra, G., & Gergely, G. (2007). 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta psychologica*, *124*(1), 60–78.

Demir, M., McNeese, N. J., Cooke, N. J., Ball, J. T., Myers, C., & Frieman, M. (2015). Synthetic teammate communication and coordination with humans. *Proceedings of the human factors and ergonomics society annual meeting*, *59*(1), 951–955.

Eelbode, T., Bertels, J., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., & Blaschko, M. B. (2020). Optimization for medical image segmentation: Theory and practice when evaluating with dice score or jaccard index. *IEEE Transactions on Medical Imaging*, *39*(11), 3679–3690.

Endsley, M. R. (2015). Situation awareness misconceptions and misunderstandings. *Journal of Cognitive Engineering and Decision Making*, *9*(1), 4–32.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frith, C., & Frith, U. (2005). Theory of mind. *Current biology*, *15*(17), R644–R645.

Gunning, D., & Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI magazine*, *40*(2), 44–58.

Gurney, N., & Pynadath, D. V. (2022). Robots with theory of mind for humans: A survey. *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 993–1000.

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, 1–8.

Ho, M. K., & Griffiths, T. L. (2022). Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Review of Control, Robotics, and Autonomous Systems*, *5*, 33–53.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. *Advances in neural information processing systems*, *29*.

Huang, L., Freeman, J., Cooke, N., Colonna-Romano, J., Wood, M. D., Buchanan, V., & Caufman, S. J. (2022). Exercises for artificial social intelligence in minecraft search and rescue for teams. *OSF doi:10.17605/osf.io/jwyvf*.

Iman, M., Rasheed, K., & Arabnia, H. R. (2022). A review of deep transfer learning and recent advancements. *arXiv preprint arXiv:2201.09679*.

Jong, N. K., Hester, T., & Stone, P. (2008). The utility of temporal abstraction in reinforcement learning. *AAMAS (1)*, 299–306.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. *International conference on machine learning*, 1885–1894.

Li, L., Walsh, T. J., & Littman, M. L. (2006). Towards a unified theory of state abstraction for mdps. *ISAIM*, *4*, 5.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, *2*(10), 749–760.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Nguyen, T. N., & Gonzalez, C. (2022). Theory of mind from observation in cognitive models and humans. *Topics in Cognitive Science*, *14*(4), 665–686.

Nowak, T., Nowicki, M. R., Ćwian, K., & Skrzypczyński, P. (2019). How to improve object detection in a driver assistance system applying explainable deep learning. *2019 IEEE Intelligent Vehicles Symposium (IV)*, 226–231.

Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., & Gombolay, M. (2021). The utility of explainable ai in ad hoc human-machine teaming. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (pp. 610–623). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/file/05d74c48b5b30514d8e9bd60320fc8f6-Paper.pdf

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. *International conference on machine learning*, 4218–4227.

Raileanu, R., Denton, E., Szlam, A., & Fergus, R. (2018). Modeling others using oneself in multi-agent reinforcement learning. *International conference on machine learning*, 4257–4266.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition*, *65*(2-3), 87–101.

Smaoui, N., & Bessassi, S. (2013). A developed system for melanoma diagnosis. *International Journal of Computer Vision and Signal Processing*, *3*(1), 10–17.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Wang, P., Wang, J., Paranamana, P., & Shafto, P. (2020). A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, *33*, 17582–17593.

Yang, S. C.-H., Folke, T., & Shafto, P. (2022). A psychological theory of explainability. *arXiv preprint arXiv:2205.08452*.

Yang, S. C.-H., & Shafto, P. (2017). Explainable artificial intelligence via bayesian teaching. *NIPS 2017 workshop on teaching machines, robots, and humans*, *2*.

Yang, S. C.-H., Vong, W. K., Sojitra, R. B., Folke, T., & Shafto, P. (2021). Mitigating belief projection in explainable artificial intelligence via bayesian teaching. *Scientific reports*, *11*(1), 1–17.

Yang, S. C.-H., Yu, Y., Wang, P., Vong, W. K., Shafto, P., et al. (2018). Optimal cooperative inference. *International Conference on Artificial Intelligence and Statistics*, 376–385.

Zheng, B., Verma, S., Zhou, J., Tsang, I., & Chen, F. (2021). Imitation learning: Progress, taxonomies and opportunities. *arXiv preprint arXiv:2106.12177*.

Zhu, H., Neubig, G., & Bisk, Y. (2021). Few-shot language coordination by modeling theory of mind. *International Conference on Machine Learning*, 12901–12911.

Zhu, X., Singla, A., Zilles, S., & Rafferty, A. N. (2018). An overview of machine teaching. *arXiv preprint arXiv:1801.05927*.

The complete Table 1 is shown here for rendering using APA7 man.

Name	Symbol	Brief description
World state	s	The current situation, including all task-relevant information that has been observed so far about the environment and the team.
Mental state	θ	The machine's inner model, which includes fine-grained decisions, intent, and goals as well as high-level constructs such as a plan; not directly observable by the human teammates.
Message	x	A collection of data that the machine provides to its human teammates to communicate its mental state.
Communication posterior	$P_T(x \mid \theta, s)$	The probability that the machine chooses message x to convey its mental state θ in the current situation s (Equation 1). The posterior's prior and likelihood are the communication prior and the inference posterior, respectively.
Communication prior	$P(x \mid s)$	The machine's probability of choosing message x in the current situation s without considering θ (Equation 1); often used to account for general information-processing constraints in humans, such as cognitive load.
Inference posterior	$P_L(\theta \mid x, s)$	The probability that the human teammate will correctly infer the θ intended to be communicated after receiving the message x given the current world state s. It is the likelihood of the communication posterior (Equation 1) as well as a posterior (Equations 2) with its own associated prior and likelihood. This inference captures the inner working of the Machine Theory of Mind.
Inference prior	$P(\theta \mid s)$	The probability that the human agent entertains mental state θ in situation s (Equation 2).
Inference likelihood	$P(x \mid \theta, s)$	The human teammate's belief that the machine would choose message x to convey a mental state θ in situation s; humans assign this probability by comparing how similar the message x is to the message that they themselves would provide (Equations 3).
Machine specification	w	The set of parameters that fully specifies the machine.

Learning mechanism	$P(w \mid x)$	The probability of updating the machine's specification to w with message x, where x is limited to have the same form as s; used in Approach 2 to construct the human teammate's the inference posterior (Equation 5).
Prediction mechanism	$P(\theta \mid w, s)$	The probability that the machine considers θ in situation s given the specification w; used together with $P(w \mid x)$ to construct the inference posterior (Equation 5).

Table 1

Glossary.

Name	Symbol	Brief description
World state	s	The current situation, including all task-relevant information that has been observed so far about the environment and the team.
Mental state	θ	The machine's inner model, which includes fine-grained decisions, intent, and goals as well as high-level constructs such as a plan; not directly observable by the human teammates.
Message	x	A collection of data that the machine provides to its human teammates to communicate its mental state.
Communication posterior	$P_T(x \mid \theta, s)$	The probability that the machine chooses message x to convey its mental state θ in the current situation s (Equation 1). The posterior's prior and likelihood are the communication prior and the inference posterior, respectively.
Communication prior	$P(x \mid s)$	The machine's probability of choosing message x in the current situation s without considering θ (Equation 1); often used to account for general information-processing constraints in humans, such as cognitive load.
Inference posterior	$P_L(\theta \mid x, s)$	The probability that the human teammate will correctly infer the θ intended to be communicated after receiving the message x given the current world state s. It is the likelihood of the communication posterior (Equation 1) as well as a posterior (Equations 2) with its own associated prior and likelihood. This inference captures the inner working of the Machine Theory of Mind.
Inference prior	$P(\theta \mid s)$	The probability that the human agent entertains mental state θ in situation s (Equation 2).
Inference likelihood	$P(x \mid \theta, s)$	The human teammate's belief that the machine would choose message x to convey a mental state θ in situation s; humans assign this probability by comparing how similar the message x is to the message that they themselves would provide (Equations 3).
Machine specification	w	The set of parameters that fully specifies the machine.
Learning mechanism	$P(w \mid x)$	The probability of updating the machine's specification to w with message x, where x is limited to have the same form as s; used in Approach 2 to construct the human teammate's the inference posterior (Equation 5).
Prediction mechanism	$P(\theta \mid w, s)$	The probability that the machine considers θ in situation s given the specification w; used together with $P(w \mid x)$ to construct the inference posterior (Equa-

Table 2

Relative strengths and weaknesses of the approaches.

Metric	Approach 1 (psychologically grounded)	Approach 2 (machine belief)	Approach 3 (domain knowledge)
Accuracy	high	low	medium
Transparency	high	low	high
Personalization	high	medium	low
Scalability	low	high	medium